TECHNOLOGY FOR DEVELOPING STATISTICAL THINKING: A PSYCHOLOGICAL PERSPECTIVE

Peter Sedlmeier
Institut für Psychologie
Technische Universität Chemnitz, Chemnitz, Germany
peter.sedlmeier@psychologie.tu-chemnitz.de

Recent years have seen tremendous advancements and innovations in technology that also can be (and have been) used for teaching statistical thinking, reasoning, and literacy. However, these modern technological tools do not automatically yield better learning results than those achieved with traditional methods of instruction. More important than technology itself is a sound theoretical basis for building an effective technological tool. It is proposed that this theoretical basis should include aspects of usability, pedagogical aspects, and also content specific aspects. A brief description of the ACT tutoring systems program illustrates what a successful combination of these three aspects could look like. Then, the importance of the often neglected content specific aspects is demonstrated with examples from our own research. It is recommended that a special emphasis should be given to a systematic and sound evaluation of such technological tools.

BACKGROUND

Recent years have seen tremendous advancements and innovations in technology that also can be (and have been) used for teaching statistical thinking, reasoning, and literacy. Such technologies have been applied for basically all kinds of statistics education, such as in online courses, in tools for data analysis, and in tutoring programs that specifically attempt to improve statistical thinking. However, so far, the huge majority of these tools seem not to have been systematically evaluated in respect to whether they are really effective. And if they have been evaluated, usually only pre-post designs without control groups have been used (e.g., Kuhn, Hoppe, & Wichmann, 2006; Mills & Raju, 2011; Raffle & Brooks, 2005). Such designs might, however, only have very low internal validity (Rosenthal & Rosnow, 1991), that is, if improvements were found, it might not be clear whether these were attributable to the new technologies or to other factors that were not controlled for. So, "…in terms of future work in this field, there is a need for well-designed studies that control for confounding variables and other challenges related to empirical research" (Mills & Raju, 2011, p. 22). Apart from methodological concerns about the evaluation studies, the results found there were quite mixed and a general superiority of teaching with the new technologies over traditional teaching methods could not be recognized (e.g., Härdle, Klinke, & Ziegenhagen, 2007; Mills & Raju, 2011). Why is this so? In this paper, I propose that a systematic improvement in statistics education by using new technologies can only be achieved if the technical aspects are well connected to theoretical aspects that are relevant to the teaching-learning process. In the next paragraph these theoretical aspects will be briefly discussed and then one of these aspects, content or task specific theories, will be illustrated in more detail.

THEORETICAL REQUIREMENTS FOR TECHNOLOGY TOOLS

Fascinating as a new technology may be, it does not automatically guarantee that users will profit from it: Users must also be able to interact appropriately with the respective tools. How to optimize this interaction can be found by trial and error, but if a tool is to be applied in the long term it is worthwhile to develop or rely on a *usability theory* that enables the tool-builders to develop adequate user interfaces. Such user interfaces are especially important in tools that are applied for teaching purposes. If the contents to be taught are complex, as is the case for statistical knowledge, it might not always be possible to build interfaces that are fully intuitively understandable. In that case it is necessary to teach the user to acquire some kind of technological literacy (Gould, 2010). How to teach that best, should be based on a *pedagogical theory*. From such a pedagogical theory one should also be able to derive directions about which kinds of teaching strategies should be used for which students under which circumstances. And finally, the contents to be taught or the task to be solved also can make a remarkable difference. In particular, it might make a huge difference how a given task is represented to the learner or recipient of statistical

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

information. The human information system is, for instance, much more sensitive to task representations that fit with the way a task naturally occurs than to other representations that also might make sense for a statistics educator (Sedlmeier, 2007). To take these specifics into account, one needs a *content specific theory*. So, to optimize instruction with the use of new technologies, it would be desirable to have at one's disposal a theory that incorporates all three of these aspects (see Figure 1).

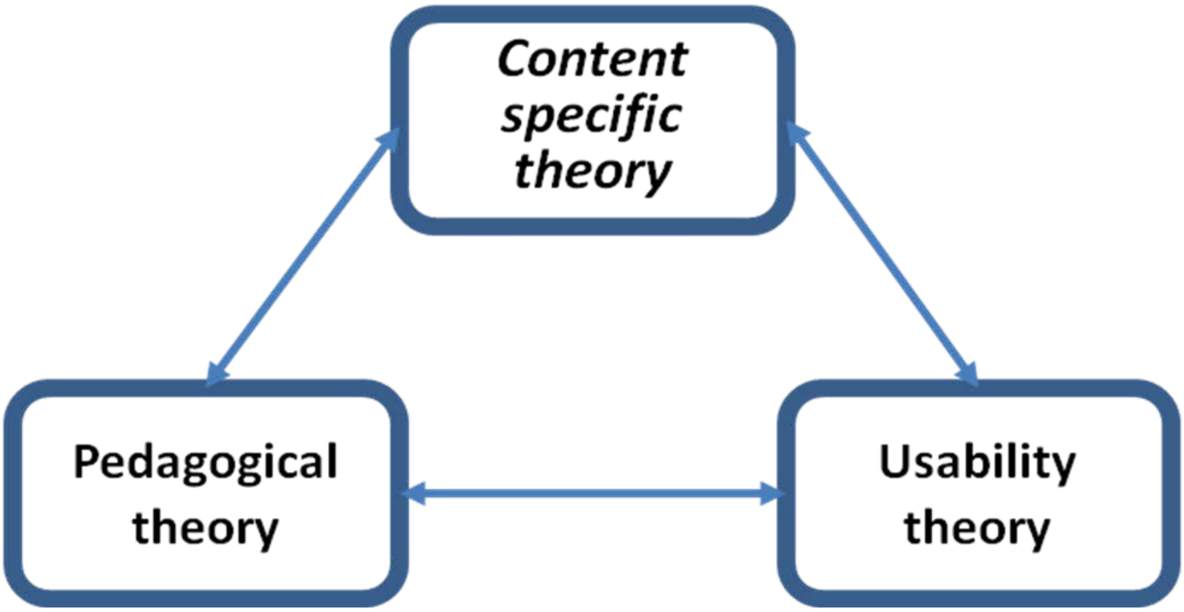

Figure 1. The three theoretical aspects that should be heeded when developing new technologies.

The three theoretical aspects, which cannot always be clearly separated because of their high degree of interdependency (as indicated by the bidirectional arrows in Figure 1) are well heeded in intelligent tutoring systems (Sedlmeier, 2001). Especially the intelligent tutoring systems based on the adaptive control of thought (ACT, later ACT-R) theory of cognition have proven very effective and useful in school curricula (Ritter, Anderson, Koedinger, & Corbett, 2007). These systems are based on a pedagogical theory, that is, a theory about how people learn and modify their knowledge, and all systems build on task analyses of the specific contents to be taught which were elaborated in empirical studies prior to constructing the tutors (Anderson, Corbett, Koedinger, & Pelletier, 1995). The ACT tutoring-system research program that has now been in existence for over 25 years, is an excellent example of how theory can guide the construction of tutoring systems, and how in turn the empirical results from evaluation studies can improve the theory (Anderson et al., 1995). In the course of the empirical application of these tutors, a kind of usability theory also evolved that yielded improvements in the pedagogical aspects of the theory and lead to specifically designed user interfaces (e.g. Koedinger & Aleven, 2007).

However, building an intelligent tutoring system is a very time consuming and complex affair, and as yet, the ACT research program does not seem to have produced a statistics tutor. But using the underlying comprehensive cognitive theory and the experience accumulated in this research group, building such a tutor would be a promising enterprise. Most attempts to using technology in statistics education are, so far, are of a much smaller scale and are usually based on only a subset of the three aspects depicted in Figure 1. Especially task specific theories usually play only a marginal role. In the following paragraph, I would like to illustrate, based on examples from our own research, why such a content specific component is very important for technological tools to be used in statistics education.

THE IMPORTANCE OF CONTENT SPECIFIC THEORIES

The usual way to convey statistical information is through use of (rational) numbers and formulas. For quite a few statistically naïve persons and students with little or no statistical knowledge, this kind of representation, instead of being helpful, tends to create math anxiety (Ashcraft, 2002). But even for persons without math anxiety, rational numbers and formulas are not always the optimal way to communicate statistical information. Fortunately, better ways exist. It makes, for instance, a decisive difference whether probabilities are presented as relative frequencies (or rational numbers) – the usual way – or as absolute frequencies as they naturally occur. This prediction can be derived from two content specific theories. One is an evolutionary account which holds that our cognitive processes are adapted to our environment and therefore, representations that catch the structure of the environment better are more intuitively understandable than others. In the case of probabilities, there is nothing in the environment that directly corresponds to relative frequencies – the only information we can perceive out there is absolute frequencies (see Cosmides & Tooby, 1996). The other content specific theoretical approach, which makes an identical prediction, stems from learning theory. It postulates that probability tasks represented in terms of relative frequencies cannot be intuitively understood by statistical novices whereas such intuitions work if the probability information is given in absolute frequencies (e.g., Sedlmeier, 2005). This approach also explains why experts in statistics are able to solve tasks like the above (using Bayes' formula) intuitively whereas novices cannot (because experts have over the years learned to use probabilities, which, after some time, are dealt with more or less intuitively).

An example task may illustrate the difference. Let us first have a look at the task in its difficult (relative frequency) format:

> A reporter for a women's monthly magazine would like to write an article about breast cancer. As a part of her research, she focuses on mammography as an indicator of breast cancer. She wonders what it really means if a woman tests positive for breast cancer during her routine mammography examination. She has the following data:
>
> - The probability that a woman who undergoes a mammography will have breast cancer is 1%.
> - If a woman undergoing a mammography has breast cancer, the probability that she will test positive is 80%.
> - If a woman undergoing a mammography does not have breast cancer, the probability that she will test positive is 10%.
>
> What is the probability that a woman who has undergone a mammography actually has breast cancer, if she tests positive?

This task (as well as similar tasks of this type) was solved correctly by less than 10% of participants, experts (medical doctors) and lay people alike, in several studies (for an overview see Sedlmeier & Gigerenzer, 2001). The usual way to solve it is to apply Bayes's formula [p(cancer) = 0.01, p(pos|cancer) = 0.8, and p(pos|no cancer) = 0.1, as given in text; p(no cancer) can be calculated as 1 – p(cancer) = 0.99]:

$$
\begin{aligned}
p(cancer \mid pos) &= \frac{p(cancer) \times p(pos \mid cancer)}{p(cancer) \times p(pos \mid cancer) + p(no\ cancer) \times p(pos \mid no\ cancer)} \\
&= \frac{.01 \times .8}{.01 \times .8 + .99 \times .1} \\
&\approx 7.5\%
\end{aligned}
$$

The low solution rates indicate that this is definitely a topic for the use of training programs, preferably ones based on the best technologies available. We devised tutoring systems that used both types of representations (see Figure 2) and compared the training results (Sedlmeier 1999; Sedlmeier & Gigerenzer, 2001). The rational numbers in the right kind of representation in Figure 2 can be inserted into Bayes' formula and the absolute frequencies (Figure 2, left) can be used to calculate the posterior probability of a woman of the respective population having cancer if she tests positive in the following way:

$$
\begin{aligned}
p(cancer|pos.) &= \frac{\#(pos. \cap cancer)}{\# pos.} \\
&= \frac{8}{8+99} \\
&= .075
\end{aligned}
$$

It turned out that the immediate training effects for both kinds of frequencies were comparably high (about 95% correct for the frequency tree and about 85% correct for the probability tree) but when participants were tested again after three months, those who had used the frequency tree in the training remained at their level whereas the solution rates for participants who had been trained with the probability tree representation diminished by about 40 percentage points. What does this example illustrate? It should illustrate the importance of having a suitable content- or task-specific theory: The seemingly equal visual representations, which would not make a difference according to a usability theory, make a remarkable difference that can be explained by a content specific theory.

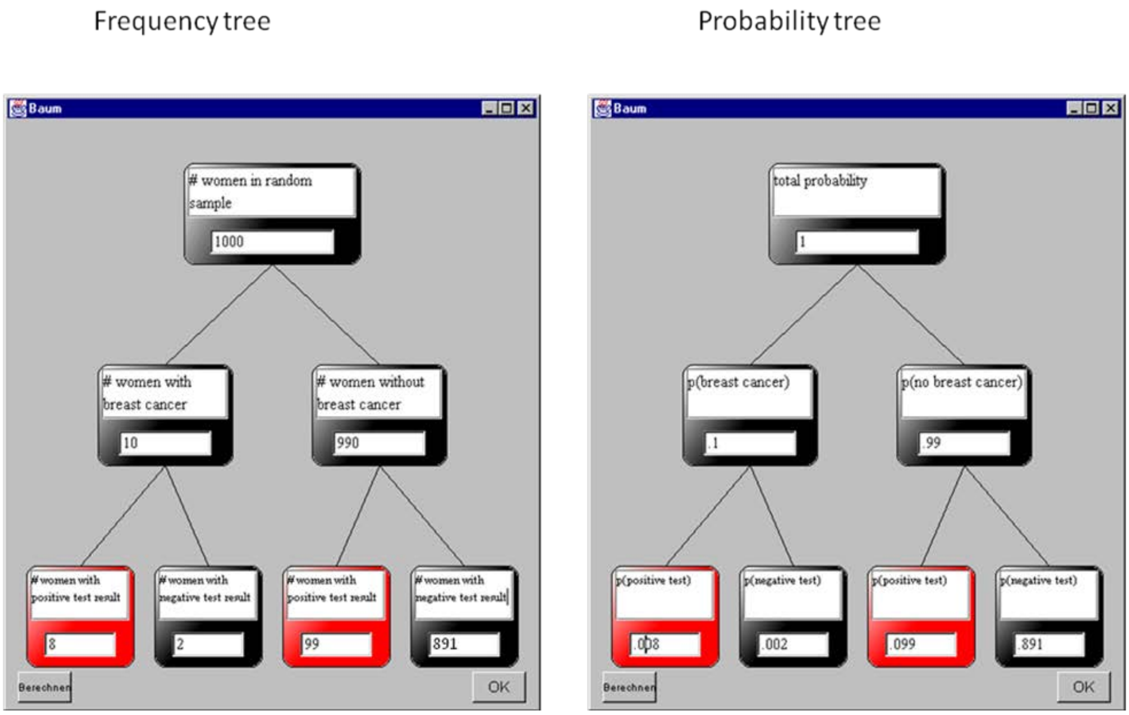

Figure 2. Two almost identical representations for a given Bayesian task that yield vastly different long-term solution rates

Using absolute frequencies is, however, not a universal solution for all kinds of tasks. It does, for instance, not work in other mathematical domains such as in integral calculus (Sedlmeier, Brockhaus, & Schwarz, 2014). And it also does not work automatically for all kind of tasks that examine the impact of sample size. Let us again illustrate this with a well known example from psychological judgmental research (Kahneman & Tversky, 1972):

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.
>
> For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys.
>
> Which hospital do you think recorded more such days? (Or was there no difference?)

Tasks of this type could be argued to use at least partially absolute frequencies – they deal with the frequency of babies and the frequency of days. However, the solution rates are usually at chance level – only about a third of participants give the correct answer (here: smaller hospital). This changes, when the task is changed slightly. If the question refers to what can be expected on a single randomly chosen day (instead a period of 1 year), the solution rates go up to an average of 75% (Sedlmeier & Gigerenzer, 1997).

Why is that so? Sedlmeier and Gigerenzer (1997) argue that we have at our disposal an intuition that conforms to the empirical law of large numbers. This intuition can be applied when comparing two frequency distributions of different size (the distributions of boys and girls on a given day) but not (at least not directly) when comparing to sampling distributions (the distributions of the percentages of baby boys over one year, for both hospitals). Illustrating both tasks with the help of computer technologies by showing an animated sampling process helps in both cases, but the difference in solution rates remains substantial (Sedlmeier, 1998). Interestingly, this difference in solution rates is predicted by associative learning theory, implemented in a neural network (Sedlmeier, 2006). The natural way samples are acquired is repetitive sampling. And if this dynamic sampling process is mimicked by a technological tool, one would expect higher solution rates, which is what happens for *frequency distribution tasks* (comparing two frequency distributions). If however, the sampling process is about some aggregated measures, such as the percentage of baby boys over the period of one year (which yields an empirical sampling distribution), then this cannot be observed directly in a natural environment and therefore cannot be expected to trigger some intuitive judgment in persons confronted with such *sampling distribution tasks*.

So what does this example illustrate? First, it should illustrate that a format that works well for one type of statistical task (e.g. a probability revision task) does not automatically also work well for another (a “sample size task”). Second, it illustrates that mimicking the actual process (by replacing a static description of the sampling process by an animation) is helpful. And third, illustrates that if tasks look quite similar at the surface levels (as frequency and sampling distribution tasks do), there might be a remarkable difference in respondents’ ability to solve them, if the task structure is in fact different. Also recommendations for teaching statistical thinking about this task would be different. Whereas for the understanding of frequency distribution tasks a good simulation of the sampling process should almost do the job of making the impact of sample size visible, sampling distribution tasks usually need some additional explanation. Taken together, the examples in this paragraph should illustrate that it is necessary for tutoring purposes, be it with or without technology, to first analyze the content or task at hand before starting extensive work on building such a tool.

CONCLUSION

This paper suggests that, to build effective technological tools that can further statistical thinking, it is necessary to have a sound theoretical basis not only for the technological part and its usability but also for the pedagogical aspects involved and for the – often neglected –content specific aspects. The importance of the latter was illustrated with examples from our own attempts to find out how teaching statistics, using technology, can be improved. It seems that, so far, there is a huge amount of work put into the development of technology in statistics education, but not so much into a systematic evaluation of the resulting tools. And it also seems that, as yet, quite a number of evaluation studies used suboptimal designs in evaluating the respective tools. If the theoretical basis for building such tools is made explicit and used for constructing the respective systems, this should yield practical benefits as well as improve the theoretical basis: fully theory guided tutoring tools can be expected to yield better learning effects and the evaluation the respective tools is again the basis for improving the theory. Such a continuing interaction between theory and practice seems to be the best way to optimize statistics education.

REFERENCES

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167-207.

Ashcraft, M.H. (2002). Math anxiety: Personal, educational, and cognitive consequences. *Directions in Psychological Science*, *11*, 181-185.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Gould, R. (2010). Statistics and the modern student. *International Statistical Review*, *79*, 297 - 315.

Härdle, W. K., Klinke, S., & Ziegenhagen, U. (2007). *On the utility of e-learning in statistics*. SFB 649 discussion paper, No. 2007,050. [Internet: http://hdl.handle.net/10419/25222].

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Koedinger, K., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review*, *19*(3), 239-264.

Kuhn, M., Hoppe, U., & Wichmann, A. (2006). Computational modelling and simulation fostering new approaches in learning probability. *Innovations in Education and Teaching International*, *43*(2), 183–194.

Mills J., D., & Raju, D. (2011). Teaching Statistics Online: A decade's review of the literature about what works *Journal of Statistics Education*, *19*(2). [Internet: www.amstat.org/publications/jse/v19n2/mills.pdf]

Raffle, H., & Brooks, G. P. (2005). Using monte carlo software to teach abstract statistical concepts: A case study. *Teaching of Psychology*, *32*(3), 193-195.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249-255.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. *Journal of Behavioral Decision Making*, *11*, 281-301.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications*. Mahwah: Lawrence Erlbaum Associates.

Sedlmeier P. (2001). Intelligent tutoring systems. In P. B. Baltes & N. J. Smelser (Eds.), *International encyclopedia of the social & behavioral sciences, Cognitive psychology and cognitive science (W. Kintsch, Section Ed.).* Amsterdam: Elsevier Science, *11*, *7674-8*.

Sedlmeier, P. (2005). From associations to intuitive judgment and decision making: Implicitly learning from experience. In T. Betsch & S. Haberstroh (Eds), *Experience based decision making* (pp. 83-99). Mahwah: Erlbaum.

Sedlmeier, P. (2006). Intuitive judgments about sample size. In K. Fiedler & P. Juslin (Eds.). *Information sampling and adaptive cognition* (pp. 53-71). Cambridge: Cambridge University Press.

Sedlmeier. P. (2007). Statistical reasoning: valid intuitions put to use. In M. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 389-419). New York: Lawrence Erlbaum Associates.

Sedlmeier, P., Brockhaus, F., & Schwarz, M. (2014). Visual integration with stock-flow models: How far can intuition carry us? In P. Bender, R. Hochmuth, P. Fischer, D. Frischemeier,& T. Wassong (Eds.), *Using tools for learning mathematics and statistics* (pp. 57-70). Berlin: Springer.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers? *Journal of Behavioral Decision Making*, *10*, 33-51.

Sedlmeier, P., & Gigerenzer, G. (2001) Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380–400.